



# Datasets: A Resource for Genomic Data from NCBI

A portal with customizable tools to access genomic sequences and their related datasets

<https://www.ncbi.nlm.nih.gov/datasets>

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

## Introduction

Advances in sequencing technology has led to dramatic increase in available genomic sequence data for a large collection of organisms. This also poses a significant challenge on how to organize and present the available datasets, and how to make these datasets readily accessible. At NCBI, genomic assemblies are organized through versioned entries in the Assembly database [1], which provides a summary of the assembly and a link to the dataset stored in the NCBI FTP site. The Assembly database also provides a download tool to allow bulk download of the retrieved set.



NCBI Datasets is a new resource that lets you easily gather assembled genomic data and their related datasets from across NCBI databases. It provides a web search page to allow customization of datasets for browsing the downloading, an API service for integration with various third party tools or workflows, and a command line tool for bulk access. This handout will address the key features of this newly released portal. Currently, the access is limited to the eukaryotic entries. Prokaryotic and viral datasets are also available for downloading, but not for online browsing.

## Getting Started

The main entry point is through the web portal, shown to the right, which provide accesses to:

- An overview of this resource given at the top (A)
- Information on programmatic access by way of command-line tool or API through linked pages (B)
- List of genome assemblies available for major taxonomic groups and well studied species (C) linking to the web search page with results limited to that group or species
- A link to a new interface for gene-specific information (D) allowing the retrieval of information/datasets for a user-specified genes
- A link to all publicly available SARS-Cov-2 datasets (E) from NCBI, and
- A collection of FAQs (F) addressing common questions over this resources

The screenshot shows the NCBI Datasets web portal. At the top is the NIH logo and the text 'National Library of Medicine National Center for Biotechnology Information'. Below this is a search bar with the text 'Search NCBI' and a 'SEARCH' button. The main heading is 'Welcome to NCBI Datasets BETA'. Below this is a paragraph: 'NCBI Datasets is an experimental resource for finding and building datasets - and we're just getting started! Our web interface allows you to download genome sequence and annotation for eukaryotic organisms and our recently added SARS-CoV-2 genome and protein datasets. ... more'. To the right of this paragraph is a table with columns: Gene ID, Symbol, Gene name, and Chromosome. The table lists several genes: 6794 (STR11), 1499 (CTNMB1), 4089 (SMAD4), 4436 (MSH2), 207 (AKT1), and 11200 (CHEK2). Below the table is a section titled 'Data tables' with the text 'Build a table of genes or transcripts and choose from a variety of custom columns.' and a 'GET STARTED' button. To the left of the 'Data tables' section is a section titled 'Programmatic access' with three sub-sections: 'Command-line' (Our Datasets command-line tool, is available for Windows, Mac, and Linux.), 'GitHub' (Explore Datasets with our Python library and Jupyter notebooks.), and 'Datasets API' (Use our RESTful APIs to add functionality to your applications.). Below the 'Programmatic access' section is a link: <https://www.ncbi.nlm.nih.gov/datasets/>. Below the link is a section titled 'Browsing genome datasets' with four sub-sections: 'Animals', 'Plants', 'Fungi', and 'Eukaryotes'. Below the 'Browsing genome datasets' section is a table with columns: Organism, Assemblies, and Download. The table lists several organisms: Homo sapiens (human), Mus musculus (house mouse), Arabidopsis thaliana (thale cress), Rattus norvegicus (Norway rat), Drosophila melanogaster (fruit fly), and Bos taurus (cattle). To the right of the 'Browsing genome datasets' section is a section titled 'Coronavirus datasets' with the text 'Download SARS-CoV-2 genome and protein sequences, annotation and a data report for all complete genomes.' and a 'GET DATA' button. Below the 'Coronavirus datasets' section is a section titled 'FAQs' with the text 'Which genomes are in NCBI Datasets?', 'What is the difference between a GenBank (GCA) and RefSeq (GCF) genome assembly?', 'What is GFF3 format?', and 'What is GBFF format?'.

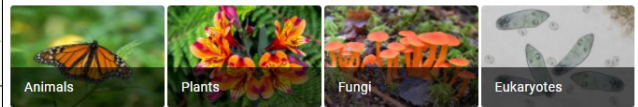
## The Interactive Search Page

The interactive search page of NCBI Datasets, accessible through taxonomic groups (A), provides functionalities that

- Displays the available genomic assemblies
- Allows filtering of available assemblies by taxonomic group
- Enables selection of assemblies of interests, and
- Allows download of selected assemblies in customized file type

Click on the image of a taxonomic group or species to access the NCBI Datasets search page. Available assemblies are filtered to that group or species. Click "X" to remove the filter.

### Browsing genome datasets



Click this to show RefSeq entries. Click again to deactivate.

Click on the Taxonomic filter to see this menu options in the popup below.

Enter text terms and select from the suggested list to jump to the desired taxonomic collection.

Retrieved assemblies are grouped by species, each under its own heading.

Key attributes of an assembly shown are: organism name, assembly name/type/ accession, assembly level, N50 stat, total size, creation date.

The assembly accession links to the record in the NCBI Assembly database.

Selecting any assembly activates the download button. Assemblies selected are indicated by the number in red.

Use checkbox to the left of an assembly to select it. Click the top-most checkbox to select all.

### Genomes – NCBI Datasets BETA

Download a genome dataset including genome, transcript and protein sequence, annotation and a data report

[NCBI Datasets](#) [Command-line tool](#) [API documentation](#)

[DOWNLOAD](#)

☐ Species

Assembly

*Abcondita terminalis*

☐ Isolate: Ate-2015

Ate (ref)

GenBank:

*Acanthaster planci*

☐ crown-of-thorns starfish

OKI-Apl

RefSeq: G

☐ crown-of-thorns starfish

OKI-Apl

GenBank:

☐ rifleman

Isolate: BGI\_N310

ASM69581.1

GenBank: GCA\_000695815.1

reference

From INSDC submitter

Scaffold

21 kb

1.036 Gbp

2014

Showing 100 of 2370 species

<< Species with assembled genomes | [Pagination button](#) >>

[SHOW MORE](#)

<https://www.ncbi.nlm.nih.gov/datasets/genomes/?txid=33208>

[DOWNLOAD](#) 13

☒ Species

*Danio rerio*

☒ zebrafish

Strain: Tuebing

☒ zebrafish

Strain: Tuebing

☒ zebrafish

Strain: CG2

☒ zebrafish

Strain: Tuebing

☒ zebrafish

Strain: Tuebing

Showing 1 of 1 species

### Download

Data from 13 assemblies

☐ Genomic sequence (FASTA)

☐ Annotated features (GFF3)

☐ Sequence and annotation (GBFF)

☐ Transcripts (FASTA)

☒ Protein (FASTA)

Your selected data and a detailed data report will be downloaded as a ZIP file.

Estimated download size is 732.42 MB

Name your file

zebrafish\_annotated\_protein.faa

[CANCEL](#) [DOWNLOAD](#)

Clicking the download button displays the download dialog box, which allows the custom selection of type of data files from the list.

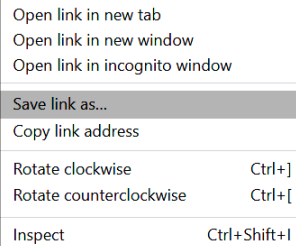
Downloaded assembly dataset will be a structured zipped archive, unzip or rehydrate with standalone datasets tool to get each assembly in its own folder or subdirectory.

Name the zip archive, else it assumes the default name of ncbi\_datasets.zip.

## The Command-line Tool

Clicking the Command-line link (A) in the web portal opens the document page (B), which provides downloading links for different platforms, and command option to get help from the console. This resource is under active development and new features, such as SARS-CoV-2 data access, will be introduced.

In this page, right click a platform link and use the "save link as ..." option to download the command-line dataset program. Most relevant for PC.



For Linux and Mac, use this curl command to save the datasets tool to the current directory.

Permission to execute needs to be set explicitly. Steps for Windows:

- Right click on the file
- Select Properties in the menu
- Click Security tab in the prompt
- Select proper individual
- Click "Edit" button to open a new prompt
- Check the checkbox in the "Allow" column for the "Read & execute" row
- Click "Apply" to activate
- Click "OK" to close the 2nd prompt
- Click "OK" to close the 1st prompt

For Mac and Linux, use this

### Get started with datasets B

The NCBI Datasets project has developed a command-line tool, **datasets**, that is used to query and download biological sequence data across all domains of life from NCBI databases.

**datasets** is currently in alpha and will be updated frequently to add new features, fix bugs, and enhance usability. Command syntax is subject to frequent changes. Please check this page often for updates.

In response to the COVID-19 pandemic, we have recently added new functionality to allow retrieval of [SARS-CoV-2 genome and protein datasets](#).

Latest release: 7.1.0 (2020-08-28)

### Download the datasets tool

Download the **datasets** command-line tool by clicking the ftp link below that matches your system:

[Linux](#)  
[Windows/64](#)  
[Mac](#)

To download **datasets** from the curl command:

#### Linux

```
$ curl -o datasets 'https://ftp.ncbi.nlm.nih.gov/pub/datasets/command-line/LATEST/linux-amd64/datasets'
```

#### Mac

```
$ curl -o datasets 'https://ftp.ncbi.nlm.nih.gov/pub/datasets/command-line/LATEST/mac/datasets'
```

### Running datasets on Linux and Mac

To enable execution of **datasets** on Linux and Mac systems, run the following command:

```
$ chmod +x datasets
```

On Mac systems, you may need to explicitly allow running **datasets** in Security & Privacy settings the first time you run the program.

### Getting help

To get help in using the tool or any of its sub

```
$ ./datasets --help
$ ./datasets <command> --help
```

### Programmatic access

Bacterial and viral data are not yet supported for online browsing. For access to data for all organisms, including bacteria and viruses, use our command line tool and RESTful APIs.

#### Command-line

Our Datasets command-line tool, is available for Windows, Mac, and Linux.

#### GitHub

Explore Datasets with our Python library and Jupyter notebooks.

#### Datasets API

Use our RESTful APIs to add functionality to your applications.

Use the **./datasets —help** to get the available top commands

Use the **./datasets top\_command\_name —help** to get detailed information for that command

As summed in the table below, the standalone datasets tool uses a two-level commands: the top commands specify the task, and the sub-commands specify the type of input and takes input and modification switches if applicable.

Top Commands	Sub-commands (with input)	Modification Switch
assembly_descriptors	assembly_accession <single accession> tax_id <single taxonomic id> tax_name <single organism name>	--refseq --limit --tax-exact-match
download	assembly <space-separated accessions> tax_id <space>  gene <space-separated gene ids>	-c <string: comma-delimited chromosome list, e.g., chr1,chrMT> -b <Boolean, include genbank flatfile> -g <Boolean, include gff3> -p <Boolean, include protein> -r <Boolean, include rna> -s <Boolean, include FASTA sequences>
gene_descriptors	gene_id <space-separated gene ids>	-
rehydrate	-	-f <string: file name>; -l <Boolean: list files only>; etc

An example download call: `./datasets download assembly GCF_000001405.39 GCF_000001635.26 -c chrX,chrMT`

## The REST API

The REST API for NCBI datasets provides a set of functions. Each function takes a clearly defined input and returns a specific type of data. The REST API landing page (right) documents these functions and demonstrates each in details through interactions. This page groups available functions into three categories based on the data returned, i.e. for Genomes (A), for annotated Genes (B), and for Viruses (C, details not shown). Asterisks marked functions have POST counterparts to work with larger batch input. Those POST functions are removed for clarity.

For each function listed, the page shows its general format and the task it performs. Clicking the “Get” button expands the display in place to show more details detailed description (D). Technical details on individual data fields under the Response column are not shown.

Clicking the “Try” button submits an actual request with the response Jason object shown in the highlighted textbox (E).

## NCBI Datasets API

v1alpha

<https://www.ncbi.nlm.nih.gov/datasets/docs/datasets-api/>

NCBI Datasets is a resource that lets you easily gather data from NCBI.

The Datasets API is still in alpha, and we're updating it often to add new functionality, iron out bugs and enhance usability. For some larger downloads, you may want to download a [dehydrated bag](#), and retrieve the individual data files at a later time.

### Genome A

Options to explore, summarize and download assembled genomes, including the associated sequence, metadata and annotation.

These genome services allow you to explore [assembled genomes](#). For a set of genomes of interest, identified by either assembly accession or taxonomic subtree, you can generate a summary, determine the package size, and download.

GET	/genome/accession/{accessions}	Get genome metadata by accession
* GET	/genome/accession/{accessions}/check	Check the validity of genome accessions
* GET	/genome/accession/{accessions}/download	Get a genome dataset by accession
* GET	/genome/accession/{accessions}/download_summary	Preview genome dataset download
GET	/genome/taxon_suggest/{taxon_query}	Get a list of taxonomy names and IDs found in the assembly dataset given a partial taxonomic name.
GET	/genome/taxon/{taxon}	Get genome metadata by taxonomic identifier
GET	/genome/taxon/{taxon}/tree	Get a taxonomic subtree by taxonomic identifier

### Gene B

Options to explore, summarize and download sequences and metadata for genes and their associated transcripts and proteins.

These gene services allow you explore [NCBI Gene](#), and for genes of interest, identified by either gene-id, symbol or RefSeq sequence accession, download a data package including metadata (tabular and YAML formats), transcript and protein sequence in FASTA format.

GET	/gene/accession/{accessions}	Get gene metadata by RefSeq Accession
GET	/gene/accession/{accessions}/download_summary	Get gene download summary by RefSeq Accession
* GET	/gene/id/{gene_ids}	Get gene metadata by GeneID
* GET	/gene/id/{gene_ids}/download	Get a gene dataset by gene ID
* GET	/gene/id/{gene_ids}/download_summary	Get gene download summary by GeneID
GET	/gene/symbol/{symbols}/taxon/{taxon}	Get gene metadata by gene symbol.
GET	/gene/symbol/{symbols}/taxon/{taxon}/download_summary	Get gene download summary by gene symbol.
GET	/gene/taxon_suggest/{taxon_query}	Get a list of taxonomy names and IDs found in the gene dataset given a partial taxonomic name.
GET	/gene/taxon/{taxon}/tree	Retrieve tax tree

### Virus C

Options to summarize and download SARS-CoV-2 and coronavirus genome and protein sequence, annotation and metadata.

These virus services allow you to retrieve coronavirus genome metadata or download genome, transcript and protein sequence in FASTA format, genome annotation in GenBank flat file (GBFF) and GenPept flat file (GPFF) formats, protein structures in protein databank format (PDB), and a YAML-formatted data report containing key host, geographic and other viral metadata.

Particular for the well-annotated SARS-CoV-2 dataset, a protein-focused dataset is available that allows you to download data for only the [SARS-CoV-2 proteins specified in your request](#).

GET /gene/accession/{accessions} Get detailed gene metadata by RefSeq Accession in a JSON output format.

**REQUEST**

**PATH PARAMETERS**

\*accessions array of string NM\_001347425.2 add-multiple RNA or Protein accessions.

**QUERY-STRING PARAMETERS**

returned\_content enum COMPLETE Default: COMPLETE Allowed: COMPLETE, IDS\_ONLY

sort\_schema.field enum SORT\_FIELD\_GENE Default: SORT\_FIELD\_GENE\_ID Allowed: SORT\_FIELD\_GENE\_ID, SORT\_FIELD\_GENE\_TYPE, SORT\_FIELD\_GENE\_SYMBOL

sort\_schema.direction enum SORT\_DIRECTION\_ Default: SORT\_DIRECTION\_ASCENDING Allowed: SORT\_DIRECTION\_ASCENDING, SORT\_DIRECTION\_DESCENDING

API Server <https://api.ncbi.nlm.nih.gov/datasets/v1alpha> Authentication No API key applied TRY

API Server <https://api.ncbi.nlm.nih.gov/datasets/v1alpha> Authentication No API key applied TRY CLEAR

Response Status: 200

RESPONSE RESPONSE HEADERS CURL

Copy

```
{
  "genes": [
    {
      "gene": {
        "gene_id": "2",
        "symbol": "A2M",
        "description": "alpha-2-macroglobulin",
        "tax_id": "9606",
        "taxname": "Homo sapiens",
        "type": "PROTEIN_CODING",
        "orientation": "minus",
        "genomic_ranges": [
```

Click RESPONSE HEADER or CURL links to toggle the text display to show the response header and the curl command (F).

RESPONSE RESPONSE HEADERS CURL

Copy

```
curl -X GET "https://api.ncbi.nlm.nih.gov/datasets/v1alpha/get"
-H "Accept: application/json"
```